INFORMATION AND CONTROL 39, 35–45 (1978)

# Finite State Predictors for Gaussian Sequences

JØRN JUSTESEN

*Institute of Circuit Theory and Telecommunication, Technical University of Denmark, Building 343, DK-2800 Lyngby, Denmark*

A finite state predictor for a Gaussian sequence with known power spectrum may be obtained by quantizing the optimal linear predictor. We analyse the structure, memory, and prediction error of the predictor by combining information theory and methods from linear filtering. The mutual information between the past and the future of the sequence provides a useful estimate of the number of bits of storage in a good predictor.

## 1. INTRODUCTION

In this paper we study finite state predictors for Gaussian sequences, and we measure their quality by the variance of the prediction error. Let $\frac{1}{2} \log M$ be the mutual information between the past and the future of the sequence. We shall demonstrate that $M^{1/2}$ is a useful estimate of the number of states that a predictor must have in order to achieve a prediction error close to that of the optimal linear predictor.

The predictors considered here are obtained by quantizing linear filters. Thus the problem may be interpreted as a special case of a more general problem concerning quantization errors in digital filtering. Since this is a nonlinear problem, the analysis must refer to a particular combination of input signal and transfer function. We have chosen the prediction problem in part because of its relative simplicity, but digital predictors are also of immediate practical interest as parts of data compression systems, channel equalizers, and digital regulators. Section 2 contains a brief statement of the linear prediction problem and a technique that has recently been suggested for analysing quantization errors.

We shall make frequent use of information theoretic arguments. Relationships between information theory, filtering, and estimation theory have often been suggested, but most results indicate only that optimal mean square estimates are also optimal in an information theoretic sense. Here we shall make use of the measure of information in a quantitative way by relating it to the number of bits stored in the predictor. In Section 3 we discuss the relationship between filtering and information theory and calculate the relevant mutual informations. In section 4 a lower bound to the prediction error for a predictor with limited information is established.



0019-9958/78/0391-0035$02.00/0

## 2. Linear Prediction

Consider a stationary zero mean Gaussian sequence $U$ with rational power spectrum

$$S(\omega) = \left| \frac{P(e^{-i\omega})}{Q(e^{-i\omega})} \right|^2 = \prod_{j=1}^{N} (1 - p_j e^{-i\omega})(1 - p_j e^{i\omega}) \Big/ \prod_{j=1}^{N} (1 - q_j e^{-i\omega})(1 - q_j e^{i\omega}).$$

All the $p_j$ and $q_j$ are assumed to be inside the unit circle. The orders of $P$ and $Q$ may be less than $N$ since some of the $p_j$ or $q_j$ could equal zero. We may think of the sequence $U$ as the output of a discrete filter with transfer function

$$H(z) = \frac{P(z^{-1})}{Q(z^{-1})} = \prod_{j=1}^{N} (1 - p_j z^{-1}) \Big/ \prod_{j=1}^{N} (1 - q_j z^{-1})$$

when the input is white Gaussian noise with variance 1. We refer to Oppenheim and Schafer (1974) for basic concepts of digital filtering.

Alternatively, the filter may be described by the difference equation

$$u(t) + \sum_{j=1}^{N} f_j u(t-j) = \sum_{j=0}^{N} g_j v(t-j), \tag{1}$$

where the $f_j$ and $g_j$ are the coefficients of $Q$ and $P$. The optimal one step prediction of $U$ can be obtained from (1) as

$$\hat{u}(t) = -\sum_{j=1}^{N} f_j u(t-j) + \sum_{j=1}^{N} g_j v(t-j), \qquad v(t) = u(t) - \hat{u}(t). \tag{2}$$

In (2), $v(t)$ is the prediction error at time $t$, and $V$ is the same white noise sequence as before. Thus the predictor has transfer function $[P(z) - Q(z)]/P(z)$ when $\hat{U}$ is taken as the output or $H^{-1}(z)$ with $V$ as the output.

A particular realization of the predictor may be described in terms of the system equations

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t), \qquad \hat{u}(t) = \mathbf{c}^t\mathbf{x}(t), \tag{3}$$

where $\mathbf{x}(t) = [X_1(t), x_2(t), \ldots, X_N(t)]$ is the state vector.

We shall analyze quantization errors in digital filters using a slightly modified version of a technique introduced by Mullis and Roberts (1976). Let

$$\mathbf{K} = E[\mathbf{x}(t)\,\mathbf{x}^t(t)]$$

be the correlation matrix of the state vector. Further define the matrix $\mathbf{W}$ such that

$$\rho^2 = E[\mathbf{e}^t(t)\,\mathbf{W}\mathbf{e}(t)]$$

is the variance of the output noise when white, but possibly mutually correlated, noise sequences $e_j(t)$ are added to $x_j(t)$. Mullis and Roberts showed that $|\mathbf{KW}|$ is invariant under a nonsingular transformation of the state vector

$$\mathbf{x}' = \mathbf{T}^{-1}\mathbf{x},$$

since

$$\mathbf{K}' = \mathbf{T}^{-1}\mathbf{K}\mathbf{T}^{-t}, \qquad \mathbf{W}' = \mathbf{T}^{t}\mathbf{W}\mathbf{T},$$

and thus

$$\mathbf{K}'\mathbf{W}' = \mathbf{T}^{-1}\mathbf{K}\mathbf{W}\mathbf{T}.$$

The determinant of $\mathbf{KW}$ is a useful measure of the number of states in a digital approximation to the linear filter.

Mullis and Roberts (1976) state that for white noise input

$$|\mathbf{K}_1\mathbf{W}| = \prod_{j,k}(p_j - q_k)^2 \Big/ \prod_{j,k}(1 - p_j p_k)^2$$

in terms of the poles and zeros of the transfer function. We shall evaluate the determinant for the predictor by considering the direct realization where $\mathbf{A}$ is the companion matrix of $P$,

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -g_N & -g_{N-1} & -g_{N-2} & \cdots & -g_2 & -g_1 \end{bmatrix}$$

If the input is white, $\mathbf{K}_1$ is the correlation matrix for $N$ successive variables from a sequence with power spectrum $|P(e^{-i\omega})|^{-2}$. Grenader and Szegö (1958) have shown that the determinant of this matrix is

$$|\mathbf{K}_1| = \left[\prod_{j,k}(1 - p_j p_k)\right]^{-1}$$

and consequently

$$|\mathbf{W}| = \prod_{j,k}(p_j - q_k)^2 \Big/ \prod_{j,k}(1 - p_j p_k).$$

When the input sequence is $U$, $\mathbf{K}$ becomes the correlation matrix of $N$ variables from a sequence with power spectrum

$$S(\omega)\,|P(e^{-i\omega})|^{-2} = |Q(e^{-i\omega})|^{-2}$$

and thus

$$|\mathbf{K}| = \left[\prod_{j,k}(1 - q_j q_k)\right]^{-1}.$$

Since $\mathbf{W}$ is independent of the input, the invariant for the predictor becomes

$$|\mathbf{KW}| = \prod_{j,k}(p_j - q_k)^2 \Big/ \prod_{j,k}(1 - p_j p_k) \prod_{j,k}(1 - q_j q_k). \tag{4}$$

If the columns of $\mathbf{T}$ are chosen as eigenvectors of $\mathbf{KW}$, both matrices are diagonlized. The eigenvalues are $l_j = k_{jj} w_{jj}$ where $k_{jj}$ is the variance of $x_j$ and $1/w_{jj}$ is the noise variance that produces an output noise of variance 1 when added to $x_j$. Thus a digital realization with this output noise requires slightly more than $\frac{1}{2} \log |\mathbf{KW}|$ bits of storage provided that all eigenvalues of $\mathbf{KW}$ are greater than 1. Notice that this condition is violated if one of the $p_j$ comes sufficiently close to one of the $q_j$. The corresponding factors in (4) do not cancel.

## 3. Prediction and Mutual Information

Let $\mathbf{S}_n$ be the correlation matrix of order $n$ for the sequence $U$, and $D_n$ the determinant of $\mathbf{S}_n$. Define

$$\mathbf{u}_n^+ = [u(0), u(1), \ldots, u(n-1)], \qquad [\mathbf{u}_j^- = u(-m), u(-m+1), \ldots, u(-1)].$$

The entropy of the vector $\mathbf{u}_n^+$ is

$$H(\mathbf{u}_n^+) = \tfrac{1}{2} \log(2\pi e)^n D_n$$

Using this expression we may write the mutual information between $\mathbf{u}_n^+$ and $\mathbf{u}_j^-$ as

$$I(\mathbf{u}_n^+, \mathbf{u}_j^-) = \tfrac{1}{2} \log(D_n D_j / D_{n+j}), \tag{5}$$

which is a special case of the relation given by Gelfand and Yaglom (1959). In particular for $n = 1$ we get

$$I[u(0), \mathbf{u}_j^-] = \tfrac{1}{2} \log(D_1 D_j / D_{j+1}), \tag{6}$$

where $D_1$ is the variance of $U$ and $D_{j+1}/D_j$ is the one step prediction error. The connection between the entropy of a Gaussian sequence and the prediction error was first noted by Elias (1951).

In Grenander and Szegö (1958) it is proved that asymptotically $D_n \simeq MG^n$, where

$$G = \exp\left\{\frac{1}{2\pi} \int_{-\pi}^{\pi} \log S(\omega)\, d\omega\right\}$$

may be interpreted as the prediction error, or in information theoretic terms, $\frac{1}{2} \log D_1/G$ is the redundancy of the sequence.

We could use (6) as a lower bound to the number of bits required in a quantized approximation to the prediction $\hat{u}(t)$ with a quantization error smaller than the optimal prediction error. However, such a bound would be a useful approximation to the memory of a digital predictor only in the simple case of a Markov sequence where

$$\hat{u}(t) = q \cdot u(t-1).$$

It was stated by Justesen (1976) that an $m$-state predictor with a prediction error close to that of the linear filter should have

$$\log m \simeq I(\mathbf{u}^-, \mathbf{u}^+),$$

where $\mathbf{u}^-$ and $\mathbf{u}^+$ indicate the semi-infinite sequences $\{u(t), t < 0\}$ and $\{u(t), t \geqslant 0\}$. Intuitively it is reasonable to describe the effect of the filter as that of collecting all relevant information about the future variables from the past observations. Here we compare this estimate to the number of bits in a quantized approximation to the linear predictor. In Section 4 it is related to a lower bound to the prediction error for an unrestricted $m$-state predictor.

Using (5) we calculate the mutual information as

$$I(\mathbf{u}^-, \mathbf{u}^+) = \lim_{n\to\infty} \tfrac{1}{2} \log(D_n^2/D_{2n}) = \tfrac{1}{2} \log M. \tag{7}$$

There has been little mention of the constant $M$ in the information theorry litterature, and we have obtained a simple expression only for rational power spectra. In Grenander and Szegö (1958), $M$ is expressed as

$$M = \exp\left\{\frac{1}{\pi} \iint_{|z|\leqslant 1} \left| \frac{H'(z)}{H(z)} \right|^2 d\sigma\right\}.$$

This integral may be calculated explicitly in terms of the poles and zeros of the transfer function

$$M = \prod_{j,k} (1 - p_j q_k)^2 \Big/ \prod_{j,k} (1 - p_j p_k) \prod_{j,k} (1 - q_j q_k). \tag{8}$$

Alternatively (8) can be derived from the result of Day (1975), who has obtained a closed form expression for $D_n$ as a function of $n$.

Gelfand and Yaglom (1959) calculated the mutual information between two random sequences by explicitly relating this problem to the orthogonal projection lemma of optimal linear estimation. Therefore, we should expect a detailed agreement between the approach based on least squares estimation and

one derived from information theory. If $\mathbf{x}(t)$ is the state vector of the linear predictor, we have

$$I[\mathbf{x}(0), \mathbf{u}^+] = I(\mathbf{u}^-, \mathbf{u}^+),$$

and this mutual information is preserved by any nonsingular transformation of $\mathbf{x}$. Even though the mutual information does not depend on the choice of state variables, one might expect a particularly convenient realization if

$$I[\mathbf{x}(0), \mathbf{u}^+] = \sum_j I[x_j(0), \mathbf{u}^+]$$

Actually we have found that this condition leads to the same realization as the diagonalization of $\mathbf{KW}$, and that

$$I[x_j(0), \mathbf{u}^+] = \tfrac{1}{2}\log(1 + k_{jj}w_{jj}) = \tfrac{1}{2}\log(1 + l_j). \tag{9}$$

Equation (9) has a simple interpretation as the mutual information between the Gaussian variable $x_j$ with variance $k_{jj}$ and $x_j + e_j$ where $e_j$ has variance $1/w_{jj}$ and would produce an output error equal to the prediction error.

In the appendix we give further details of this relationship, and we prove that the matrices associated with the transfer functions

$$H(z) = Q(z) = \prod_j (1 - q_j z^{-1}) \qquad \text{and} \qquad \tilde{H}(z) = \tilde{Q}(z) = \prod_j (z^{-1} - q_j)$$

satisfy

$$\tilde{\mathbf{K}}\tilde{\mathbf{W}} = \mathbf{KW} + \mathbf{I}.$$

Thus

$$I(\mathbf{u}^-, \mathbf{u}^+) = \tfrac{1}{2}\log |\, \tilde{\mathbf{K}}\tilde{\mathbf{W}} \,|.$$

If $Q$ is replaced by the reversed polynomial $\tilde{Q}$ in Eq. (4) we obtain the same expression as in (8).

From Eq. (9) we conclude that approximately $\frac{1}{2}\log(1 + l_j)$ bits should be assigned to $x_j$. This should be compared to $\max\{0, \frac{1}{2}\log l_j\}$ which was derived in Section 2. There is little difference between the two estimates, but it is an advantage of equations (8) and (9) that they are valid when pairs of poles and zeros cancel.

Finally we note that if we minimize the output noise variance $\rho^2$ for fixed entropy of the error vector $\mathbf{e} = [e_1, e_2, \ldots, e_N]$, it is easy to prove that the noise correlation matrix should be proportional to $\mathbf{W}^{-1}$. Thus for a diagonal $\mathbf{W}$, we should use independent noise with variances $1/w_{jj}$.

## 4. Predictors with Limited Information

In this section we shall derive a lower bound to the variance of the prediction error when the information about the future is restricted

$$I(\mathbf{x}, \mathbf{u}^+) \leqslant \log m.$$

Clearly this is also a lower bound to the prediction error for an $m$-state predictor.

A predictor must work in real time, and thus no coding of state vectors associated with different instances of time is possible. However, in principle we could justify the use of information quantities by considering the simultaneous coding of an ensemble of predictors working on independent inputs with identical power spectra.

The sequential operation of the predictor makes it difficult in general to obtain a precise relationship between the number of states and the prediction error. Even if we assume a linear model and use a white noise approximation to the quantization error, the analysis is difficult, because the best transfer function will differ from that of the optimal filter.

The mutual information may be expressed as the sum

$$I(\mathbf{u}^-, \mathbf{u}^+) = I[\mathbf{u}^-, u(0)] + I[\mathbf{u}^-, u(1) \mid u(0)] + I[\mathbf{u}^-, u(2) \mid u(0)\, u(1)] + \cdots.$$

Substituting

$$I[\mathbf{u}^-, u(j) \mid u(0)\, u(1) \cdots u(j-1)] = \tfrac{1}{2} \log(GD_{j+1}/D_j)$$

with $G = 1$ into this expression we obtain

$$I(\mathbf{u}^-, \mathbf{u}^+) = \tfrac{1}{2} \log D_1 + \tfrac{1}{2} \log(D_2/D_1) + \cdots = \tfrac{1}{2} \log M. \tag{10}$$

Similarly

$$\begin{aligned} I[\mathbf{u}^- u(0)\, u(1)...,u(j)] &= I[u(j-1),u(j)] + I[u(j-2), u(j) \mid u(j-1)] \\ &\quad + \cdots + I[\mathbf{u}^-, u(j) \mid u(0)\, u(1),..., u(j+1)] \end{aligned} \tag{11}$$

We interpret (10) and (11) as indicating that in an optimal linear predictor the state vector $\mathbf{x}(0)$ contains $\frac{1}{2} \log(D_{j+1}/D_j)$ bits of information about $u(j)$ which cannot be obtained from $u(0), u(1),..., u(j-1)$. Thus if this information is discarded, the prediction error must increase accordingly. For a particular error variance $d$ the greatest possible reduction of the mutual information is obtained if each term in (10) is reduced by $\frac{1}{2} \log d$ or discarded if $D_{j+1}/D_j < d$.

We may gain a better understanding of the relationship between the mutual information and the prediction error by considering filters with finite impulse responses satisfying the Wiener–Hopf equation

$$\mathbf{S}_n \mathbf{h}_n = (s_1, s_2, ..., s_n)^t, \tag{12}$$

where $s_j = E[u(t)\, u(t+j)]$. This filter generates the optimal linear estimate of $u(0)$ from $\mathbf{u}_n^-$. Further it contains all information between $u(1)$ and $\mathbf{u}_{n-1}$ given $u(0)$, whereas $u(-n)$ has no effect on $\hat{u}(1)$. In general the useful amount of information is

$$\sum_{j=1}^{n} I[\mathbf{u}_{n-j+1}^- u(j) \mid u(0)\, u(1) \cdots u(j-1)] = \sum_{j=1}^{n} \tfrac{1}{2} \log \frac{D_j D_n}{D_{j-1} D_{n+1}} = \tfrac{1}{2} \log \frac{D_n^{n+1}}{D_{n+1}^n} \quad (13)$$

In (13) each term of (10) has been reduced by the amount $\frac{1}{2}\log(D_{n+1}/D_n)$, which equals the first term that has been descarded. $D_{n+1}/D_n$ is also the pre-prediction error of the $n$th-order predictor.

We conclude that in order to achieve a prediction error with variance $\leqslant d$, the state vector must satisfy

$$I(\mathbf{x}, \mathbf{u}^+) \geqslant \tfrac{1}{2} \log D_n - \frac{n}{2} \log d, \qquad \frac{D_{n+1}}{D_n} \leqslant d \leqslant \frac{D_n}{D_{n-1}}. \quad (14)$$

In the lower bound (14) most of the reduction of the mutual information is a result of the truncation of the impulse response. In general a good finite state predictor will be an approximation to a linear filter which is less sensitive to quantization noise than the optimal predictor. Thus poles will move away from the unit circle or infinite impulse response may be replaced by a finite one.

If the finite impulse response filter is realized as a system with diagonal **KW** matrix, the predictor has certain desirable properties. If $l_j$ is an eigenvalue of **KW**, the mutual information may be written as a sum similar to the expression in Eq. (9),

$$I(\mathbf{x}, \mathbf{u}^+) = \sum_j I(x_j, \mathbf{u}^+) = \sum_j \tfrac{1}{2} \log(1 + l_j/d). \quad (15)$$

This relation may be explained by noting that a filter satisfying (12) is an optimal predictor for a sequence which is generated by passing white noise of variance $d$ through a filter with transfer function

$$Q'(z)^{-1} = \left[\prod_j (1 - q_j' z^{-1})\right]^{-1}.$$

Another desirable property of the finite impulse response filter is that the power spectrum of the quantization noise is white. To prove this assume that the state variables have been scaled to make **W** a unit matrix. If the impulse response from $x_j$ to the output is written as a vector

$$\mathbf{r}_j = [r_j(1), r_j(2), \ldots, r_j(n)],$$

we have $\mathbf{r}_j \cdot \mathbf{r}_k = w_{jk}$, which shows that the $\mathbf{r}_j$ are orthonormal. Thus the $n$

vectors $[r_1(t), r_2(t),..., r_n(t)]$ for $1 \leqslant t \leqslant n$ are also orthonormal, and the correlation function of the output noise sequence

$$E[e(0)\, e(t)] = E\left[\left(\sum_j \sum_k e_k(-j)\, r_k(j)\right)\left(\sum_j \sum_k e_k(t-j)\, r_k(j)\right)\right]$$

$$= \sum_j \sum_k r_k(j)\, r_k(j+t)$$

equals zero if $t \neq 0$.

## 5. Conclusion

A finite state predictor for a Gaussian sequence may be anlyzed as a linear system with noisy state variables.

Equation (14) relates the prediction error $d$ to the mutual information between the state and the future of the sequence $I(\mathbf{x}, \mathbf{u}^+)$. The total information $\frac{1}{2} \log M$ may be reduced by approximately $\frac{1}{2} \log d$ times the duration of the impulse response of the optimal filter. If a prediction error variance $\sim 2$ is acceptable, $\frac{1}{2} \log M$ is a good approximation to the necessary information unless one or more poles of the transfer function are very close to the unit circle. Thus Eq. (8) indicates how the zeros of the transfer function influence the number of states and how poles and zeros interact, but the effect of poles close to the unit circle may be overestimated.

If $y$ and $\hat{y}$ are Gaussian with variances $D$ and $D - 1$ and mutual information $\frac{1}{2} \log D$, we can use rate-distortion theory to give a more precise bound to the error $E[(y - y')^2]$ when $y'$ is an $m$-bit representation of $\hat{y}$. It is easy to calculate that the error satisfies

$$d \geqslant 1 + 2^{-2m}(D - 1).$$

Thus if $D$ is not too small we have $d \sim 2$ for $m = \frac{1}{2} \log D$. A similar analysis for an $N$th-order linear filter is difficult. However, the error variance is between 1 and $N$ in this case. Similarly it may be concluded from equation (9) that if $M$ is not very small, the quantization noise has variance between 1 and $N$ times the prediction error of the noisefree system.

## Appendix

Let $\mathbf{S}_N$ be the correlation matrix for a sequence with power spectrum $S(\omega) = [Q(e^{i\omega})\, Q(e^{-i\omega})]^{-1}$. In this case

$$I(\mathbf{u}^-, \mathbf{u}^+) = I(\mathbf{u}_N{}^-, \mathbf{u}_N{}^+),$$

and thus all relevant information can be obtained from the matrix

$$\mathbf{S}_{2N} = \begin{bmatrix} \mathbf{S}_N & \mathbf{S}_N\mathbf{A}^{tN} \\ \mathbf{A}^N\mathbf{S}^N & \mathbf{S}_N \end{bmatrix},$$

where $\mathbf{A}$ is the companion matrix of the polynomial $Q(z)$. If we consider the direct realization of the predictor with transfer function $Q(z)$, $\mathbf{S}_N$ is the correlation matrix of the state vector, and the output error weight matrix is

$$\mathbf{W} = \begin{bmatrix} f_N & 0 & \cdots & 0 \\ f_{N-1} & f_N & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ f_1 & f_2 & \cdots & f_N \end{bmatrix} \begin{bmatrix} f_N & f_{N-1} & \cdots & f_1 \\ 0 & f_N & \cdots & f_2 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & f_N \end{bmatrix}.$$

Let $\mathbf{T}$ satisfy

$$\mathbf{T}^{-1}\mathbf{S}_N\mathbf{T}^{-t} = \mathbf{I}, \qquad \mathbf{T}\mathbf{W}\mathbf{T}^t = \operatorname{diag}\{l_1, l_2, \ldots, l_N\},$$

where $\mathbf{I}$ is a unit matrix. Then

$$\mathbf{S}'_{2N} = \begin{bmatrix} \mathbf{T}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{S}_N & \mathbf{S}_N\mathbf{A}^{tN} \\ \mathbf{A}^N\mathbf{S}_N & \mathbf{S}_N \end{bmatrix} \begin{bmatrix} \mathbf{T}^{-t} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-t} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{I} & \mathbf{T}^{-1}\mathbf{S}_N\mathbf{A}^{tN}\mathbf{T}^{-t} \\ \mathbf{T}^{-1}\mathbf{A}^N\mathbf{S}_N\mathbf{T}^{-t} & \mathbf{I} \end{bmatrix}.$$

Now

$$I(\mathbf{u}_N^-, \mathbf{u}_N^+) = |\, \mathbf{S}'_{2N} \,|^{-1} = |\, \mathbf{I} - \mathbf{T}^{-1}\mathbf{A}^N\mathbf{S}_N\mathbf{A}^{tN}\mathbf{T}^{-t} \,|^{-1},$$

and the condition

$$I(\mathbf{u}_N^-, \mathbf{u}_N^+) = \sum_j I(x_j, \mathbf{u}_N^+)$$

is satisfied if $\mathbf{I} - \mathbf{T}^{-1}\mathbf{A}^N\mathbf{S}_N\mathbf{A}^{tN}\mathbf{T}^{-t}$ is diagonal. $\mathbf{S}_N$ satisfies the relation

$$\mathbf{S}_N = \mathbf{A}\mathbf{S}_N\mathbf{A}^t + \mathbf{b}\mathbf{b}^t, \qquad \mathbf{b}^t = (0, 0, \ldots, 1). \tag{16}$$

By repeated application of (16) we get

$$\mathbf{A}^N\mathbf{S}_N\mathbf{A}^{tN} = \mathbf{S}_N - \mathbf{B}^{-1}\mathbf{B}^{-t},$$

where

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_1 & 1 & 0 & \cdots & 0 \\ f_2 & f_1 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ f_{N-1} & f_{N-2} & f_{N-3} & \cdots & 1 \end{bmatrix},$$

and thus

$$\mathbf{I} - \mathbf{T}^{-1}\mathbf{A}^N\mathbf{S}_N\mathbf{A}^{tN}\mathbf{T}^{-t} = \mathbf{T}^{-1}\mathbf{B}^{-1}\mathbf{B}^{-t}\mathbf{T}^{-t}.$$

The matrix $\tilde{\mathbf{W}} = \mathbf{B}\mathbf{B}^t$ is the output error weight matrix for a filter with transfer function $\tilde{Q}(z)$ which is obtained by reversing the coefficients of $Q(z)$.
The inverse of the Toeplitz matrix $\mathbf{S}_N$ may be expressed as

$$\mathbf{S}_N^{-1} = \tilde{\mathbf{W}} - \mathbf{W},$$

and thus

$$\mathbf{S}_N\tilde{\mathbf{W}} = \mathbf{S}_N\mathbf{W} + \mathbf{I}.$$

We conclude that $\mathbf{W}$ and $\tilde{\mathbf{W}}$ are diagonalized by the same transformation, and that this transformation also separates the mutual information in the state variables. Since

$$\mathbf{T}^t\tilde{\mathbf{W}}\mathbf{T} = \text{diag}\{1 + l_1, 1 + l_2, \ldots, 1 + l_N\}$$

we have

$$I(x_j, \mathbf{u}_N^+) = \tfrac{1}{2}\log(1 + l_j).$$

## Acknowledgment

The author would to thank Tom Høholt for several helpful discussions concerning the subject matter of this paper.

Received: April 7, 1977; revised: December 16, 1977

## References

Day, M. (1975), Toeplitz matrices generated by the Laurent series expension of an arbitrary rational function, *Trans. Amer. Math. Soc.* **206**, 224–245.

Elias, P. (1951), A note on autocorrelation and entropy, *Proc. IRE* **39**, 839.

Gelfand, I. M. and Yaglom, A. M. (1959), Calculation of the amount of information about a random function contained in another such function, *Amer. Math. Soc. Transl.*

Grenander, U., and Szegö, G. (1958), "Toeplitz Forms and Their Applications," Univ. of California Press, Berkeley.

Justesen, J. (1976), The necessary memory of certain digital filters, IEEE Int. Symp. Inform. Theory, Ronneby, Sweden.

Mullis, C. T., and Roberts, R. A. (1976), Synthesis of minimum roundoff noise fixed point digital filters, *IEEE Trans. on Circuit and Systems* **CAS-23**, 551–562.

Oppenheim, A. V., and Schafer, R. W. (1974), "Digital Signal Processing," Prentice–Hall, Englewood Cliffs, N.J.